Building a Data Warehouse
With Examples in SQL Server

Vincent Rainardi
For my lovely wife, Ivana.
# Contents

About the Author ................................................................. xiii
Preface .............................................................................. xv

## Chapter 1  Introduction to Data Warehousing ................. 1

What Is a Data Warehouse? .................................................. 1
  Retrieves Data .............................................................. 4
  Consolidates Data ......................................................... 5
  Periodically ................................................................. 6
  Dimensional Data Store ................................................. 7
  Normalized Data Store ................................................. 8
  History .................................................................. 10
  Query .................................................................. 11
  Business Intelligence .................................................... 12
  Other Analytical Activities ........................................... 14
  Updated in Batches ....................................................... 15
  Other Definitions ......................................................... 16

Data Warehousing Today ................................................. 17
  Business Intelligence .................................................... 17
  Customer Relationship Management .......................... 18
  Data Mining ............................................................... 19
  Master Data Management (MDM) ............................... 20
  Customer Data Integration .......................................... 23

Future Trends in Data Warehousing ............................... 24
  Unstructured Data ....................................................... 24
  Search .................................................................. 25
  Service-Oriented Architecture (SOA) ......................... 26
  Real-Time Data Warehouse ........................................ 27

Summary ...................................................................... 27
CHAPTER 2  Data Warehouse Architecture ............................................. 29
  Data Flow Architecture .......................................................... 29
    Single DDS ................................................................. 33
    NDS + DDS ............................................................... 35
    ODS + DDS ............................................................... 38
    Federated Data Warehouse .................................................. 39
  System Architecture ............................................................. 42
  Case Study .............................................................................. 44
  Summary .................................................................................. 47

CHAPTER 3  Data Warehouse Development Methodology ...................... 49
  Waterfall Methodology ............................................................ 49
  Iterative Methodology ............................................................. 54
  Summary .................................................................................. 59

CHAPTER 4  Functional and Nonfunctional Requirements ..................... 61
  Identifying Business Areas ....................................................... 61
  Understanding Business Operations ......................................... 62
  Defining Functional Requirements .......................................... 63
  Defining Nonfunctional Requirements ..................................... 65
  Conducting a Data Feasibility Study ......................................... 67
  Summary .................................................................................. 70

CHAPTER 5  Data Modeling ................................................................. 71
  Designing the Dimensional Data Store ...................................... 71
  Dimension Tables ..................................................................... 76
  Date Dimension ...................................................................... 77
  Slowly Changing Dimension .................................................... 80
  Product, Customer, and Store Dimensions ............................... 83
  Subscription Sales Data Mart .................................................... 89
  Supplier Performance Data Mart .............................................. 94
  CRM Data Marts ..................................................................... 96
  Data Hierarchy ....................................................................... 101
  Source System Mapping .......................................................... 102
  Designing the Normalized Data Store ...................................... 106
  Summary .................................................................................. 111
CHAPTER 6  Physical Database Design ............................. 113
    Hardware Platform ......................................... 113
    Storage Considerations ................................... 120
    Configuring Databases ..................................... 123
    Creating DDS Database Structure ......................... 128
    Creating the Normalized Data Store ...................... 139
    Using Views .................................................. 157
    Summary Tables ............................................. 161
    Partitioning .................................................. 162
    Indexes ........................................................ 166
    Summary ........................................................ 171

CHAPTER 7  Data Extraction .......................................... 173
    Introduction to ETL ......................................... 173
    ETL Approaches and Architecture ......................... 174
    General Considerations .................................... 177
    Extracting Relational Databases ......................... 180
       Whole Table Every Time ................................. 180
       Incremental Extract .................................... 181
       Fixed Range .............................................. 185
       Related Tables .......................................... 186
       Testing Data Leaks ...................................... 187
    Extracting File Systems .................................... 187
    Extracting Other Source Types ......................... 190
    Extracting Data Using SSIS ............................... 191
    Memorizing the Last Extraction Timestamp ............... 200
    Extracting from Files ...................................... 208
    Summary ........................................................ 214

CHAPTER 8  Populating the Data Warehouse ...................... 215
    Stage Loading ............................................... 216
    Data Firewall ............................................... 218
    Populating NDS ............................................. 219
    Using SSIS to Populate NDS ................................ 228
    Upsert Using SQL and Lookup .............................. 235
    Normalization ............................................... 242
    Practical Tips on SSIS .................................... 249
Managing Reports ................................................................. 370
  Managing Report Security .................................................. 370
  Managing Report Subscriptions ........................................... 372
  Managing Report Execution ............................................... 374
Summary ................................................................. 375

CHAPTER 12 Multidimensional Database ........................................ 377
  What a Multidimensional Database Is ........................................ 377
  Online Analytical Processing .................................................. 380
  Creating a Multidimensional Database ...................................... 381
  Processing a Multidimensional Database .................................... 388
  Querying a Multidimensional Database ....................................... 394
  Administering a Multidimensional Database ................................ 396
    Multidimensional Database Security ......................................... 397
    Processing Cubes ............................................................ 399
    Backup and Restore .......................................................... 405
Summary ................................................................. 409

CHAPTER 13 Using Data Warehouse for Business Intelligence ................. 411
  Business Intelligence Reports ................................................ 412
  Business Intelligence Analytics ............................................... 413
  Business Intelligence Data Mining .......................................... 416
  Business Intelligence Dashboards ............................................ 432
  Business Intelligence Alerts .................................................. 437
  Business Intelligence Portal .................................................... 438
Summary ................................................................. 439

CHAPTER 14 Using Data Warehouse for Customer Relationship Management .. 441
  Single Customer View .......................................................... 442
  Campaign Segmentation ....................................................... 447
  Permission Management ....................................................... 450
  Delivery and Response Data .................................................. 454
  Customer Analysis .............................................................. 460
  Customer Support ............................................................. 463
  Personalization ................................................................. 464
  Customer Loyalty Scheme ..................................................... 465
Summary ................................................................. 466
CHAPTER 15  Other Data Warehouse Usage ................................................. 467
  Customer Data Integration ................................................................. 467
  Unstructured Data ............................................................................ 470
  Search in Data Warehousing ............................................................... 474
  Summary ............................................................................................. 476

CHAPTER 16  Testing Your Data Warehouse ............................................. 477
  Data Warehouse ETL Testing ............................................................... 478
  Functional Testing ............................................................................. 480
  Performance Testing ......................................................................... 482
  Security Testing ................................................................................ 485
  User Acceptance Testing ................................................................. 486
  End-to-End Testing ........................................................................... 487
  Migrating to Production ................................................................. 487
  Summary ............................................................................................. 489

CHAPTER 17  Data Warehouse Administration .......................................... 491
  Monitoring Data Warehouse ETL ....................................................... 492
  Monitoring Data Quality ................................................................... 495
  Managing Security ............................................................................ 498
  Managing Databases ........................................................................ 499
  Making Schema Changes .................................................................. 501
  Updating Applications ...................................................................... 503
  Summary ............................................................................................. 503

APPENDIX  Normalization Rules ........................................................... 505

INDEX ..................................................................................................... 509
About the Author

VINCENT RAINARDI is a data warehouse architect and developer with more than 12 years of experience in IT. He started working with data warehousing in 1996 when he was working for Accenture. He has been working with Microsoft SQL Server since 2000. He worked for Lastminute.com (part of the Travelocity group) until October 2007. He now works as a data warehousing consultant in London specializing in SQL Server. He is a member of The Data Warehousing Institute (TDWI) and regularly writes data warehousing articles for SQLServerCentral.com.
Friends and colleagues who want to start learning data warehousing sometimes ask me to recommend a practical book about the subject matter. They are not new to the database world; most of them are either DBAs or developers/consultants, but they have never built a data warehouse. They want a book that is practical and aimed at beginners, one that contains all the basic essentials. There are many data warehousing books on the market, but they usually cover a specialized topic such as clickstream, ETL, dimensional modeling, data mining, OLAP or project management and therefore a beginner would need to buy five to six books to understand the complete spectrum of data warehousing. Other books cover multiple aspects, but they are not as practical as they need to be, targeting executives and project managers instead of DBAs and developers.

Because of that void, I took a pen (well, a laptop really) and spent a whole year writing in order to provide a practical, down-to-earth book containing all the essential subjects of building a data warehouse, with many examples and illustrations from projects that are easy to understand. The book can be used to build your first data warehouse straightaway; it covers all aspects of data warehousing, including approach, architecture, data modeling, ETL, data quality, and OLAP. I also describe some practical issues that I have encountered in my experience—issues that you’ll also likely encounter in your first data warehousing project—along with the solutions.

It is not possible to show examples, code, and illustrations for all the different database platforms, so I had to choose a specific platform. Oracle and SQL Server provide complete end-to-end solutions including the database, ETL, reporting, and OLAP, and after discussions with my editor, we decided to base the examples on SQL Server 2005, while also making them applicable to future versions of SQL Server such as 2008. I apologize in advance that the examples do not run on SQL Server 2000; there is just too big a gap in terms of data warehousing facilities, such as SSIS, between 2000 and 2005.

Throughout this book, together we will be designing and building a data warehouse for a case study called Amadeus Entertainment. A data warehouse consist of many parts, such as the data model, physical databases, ETL, data quality, metadata, cube, application, and so on. In each chapter, I will cover each part one by one. I will cover the theory related to that part, and then I will show how to build that part for the case study. Specifically, Chapter 1 introduces what a data warehouse is and what the benefits are. In Chapters 2–6, we will design the architecture, define the requirements, and create the data model and physical databases, including the SQL Server configuration. In Chapters 7–10 we will populate the data stores using SSIS, as well as discuss data quality and metadata. Chapters 11–12 are about getting the data out by using Reporting Services and Analysis Services cubes. In Chapters 13–15, I’ll discuss the application of data warehouse for BI and CRM as well as CDI, unstructured data, and search. I close the book with testing and administering a data warehouse in Chapters 16–17.
The supplementary material (available on the book’s download page on the Apress website, http://www.apress.com) provides all the necessary material to build the data warehouse for the case study. Specifically, it contains the following folders:

**Scripts**: Contains the scripts to build the source system and the data warehouse, as explained in Chapters 5 and 6.

**Source system**: Contains the source system databases required to build the data warehouse for the case study in Chapters 7 and 8.

**ETL**: Contains the SSIS packages to import data into the data warehouse. Chapters 7 and 8 explain how to build these packages.

**Report**: Contains the SSRS reports explained in Chapter 11.

**Cubes**: Contains the SSAS projects explained in Chapter 12.

**Data**: Contains the backup of data warehouse database (the DDS) and Analysis Services cube, which are used for reporting, OLAP, BI, and data mining in Chapters 11, 12, and 13.